Towards Integrated 3D Reconstruction and Semantic Interpretation of Urban Scenes

MAROŠ BLAHÁ[1], CHRISTOPH VOGEL[2], AUDREY RICHARD[1], JAN D. WEGNER[1], THOMAS POCK[2] & KONRAD SCHINDLER[1]

We describe recent progress made towards automated extraction of semantically annotated 3D city models from aerial imagery. The idea of semantic 3D reconstruction is to reconstruct the 3D geometry of an observed scene while at the same time also interpreting the scene in terms of semantic object classes (such as buildings, vegetation etc.) - similar to a human operator, who also interprets the image content while making measurements. The advantage of jointly reasoning about shape and object class is that one can exploit class-specific a-priori knowledge about the geometry: on the one hand the type of object provides information about its shape, e.g. walls are likely to be vertical, whereas streets are not; on the other hand, 3D geometry is also an important cue for classification, e.g. in our example vertical surfaces are more likely to be walls than streets. Recent work has developed computational models that allow one to jointly infer geometry and class. For simple priors, such as preferred surface orientations, these models even have favourable mathematical properties like convexity of the optimisation. However, they rely on a dense, volumetric discretisation of 3D space, therefore the computation is memory-hungry and slow. We have developed an adaptive, hierarchical formulation of semantic 3D reconstruction, which makes it possible to process scenes of much larger, realistic size. The intuition is that both high spatial resolution and high numerical precision are only required in regions that are likely to contain a surface. Technically, our scheme amounts to repeatedly solving a constrained, convex optimisation problem, while iteratively removing low-confidence constraints. In our experiments the adaptive reconstruction incurs no loss in accuracy, but offers up to 98% lower memory consumption and up to 95% shorter computation time.

1 Introduction

Photogrammetric mapping encompasses two visual tasks: to *geometrically reconstruct* the observed scene in 3D, and to *semantically interpret* the data. These two tasks are not independent of each other, and human operators naturally solve them in conjunction, utilising their a-priori knowledge about the shape of different types of objects like buildings, water surfaces etc. On the contrary, computer systems for automated photogrammetric analysis treat the two tasks in isolation and sequentially. Either geometry is reconstructed as a generic surface model, which can then serve as input to extract objects like buildings, the ground (DTM) etc. Or the images are first interpreted to detect objects, which can then be individually reconstructed.

In which order should one go about these two steps? Arguably one should address them jointly, because a-priori knowledge about the world acts in both ways. A building probably has vertical walls; but conversely, a concrete-grey structure is much more likely to be a building wall if it is

[1] ETH Zürich, Photogrammetry and Remote Sensing, Stefano-Franscini-Platz 5, CH-8093 Zürich, E-Mail: [maros.blaha, audrey.richard, jan.wegner, schindler]@geod.baug.ethz.ch
[2] TU Graz, Computer Graphics and Vision, Inffeldgasse 16, A-8010 Graz, E-Mail: [christoph.vogel, pock]@icg.tugraz.at

vertical. However, treating both the 3D geometry and the semantic class labels as unknowns and inferring them together is technically challenging. Only recently models have emerged that capture the problem in a principled manner (HÄNE et al. 2013, SAVINOV et al. 2015), but they are computationally expensive and therefore limited to small regions of interest and/or few images.

Here, we describe a way to make semantic 3D reconstruction a lot more efficient, such that it scales up to larger regions and image sets. The application we are interested in is the generation of interpreted 3D city models from (nadir and oblique) aerial images. We follow the pioneering work of (HÄNE et al. 2013), where semantic reconstruction is formulated as a multi-class labelling problem on a voxel grid. Like in volumetric surface reconstruction (CURLESS & LEVOY 1996) the 3D space is discretised into voxels. But rather than only being labelled as *freespace* or *solid*, voxels can take on multiple labels like *freespace, building, ground, vegetation*, etc.

Our work builds on the elementary insight that instead of a regular voxel grid the labelling can be done with variable volumetric resolution. Large parts of the volume – in particular freespace, but also the inside of buildings, areas under the ground, etc. – only need to be modelled at coarse resolution. A fine discretisation and high numerical precision are only required near the boundary surfaces. We start from a coarse voxel grid and adaptively refine the reconstruction only near (predicted) label transitions. On the one hand this reduces memory consumption, so that at a given target resolution one can reconstruct larger scenes. On the other hand, it also speeds up the computation, because after every refinement an approximate solution is already available from the previous level, like in a multi-scale pyramid. In our experiments we observe up to 40× lower memory footprint and 22× shorter runtime, without any loss of quality.

2 Related Work

Automated 3D city reconstruction is a classical problem of photogrammetry. For purely geometric surface reconstruction, volumetric representations are rare and explicit surface representations are preferred. In this context (LAFARGE & MALLET 2012) have already proposed to first estimate semantic labels for 3D points and use those to support geometric reconstruction (but not vice versa). Early attempts to merge geometric and semantic reconstruction into a one-shot process started with depth maps (LADICKÝ et al. 2010), later work moved to volumetric representations (HÄNE et al. 2013, BAO et al. 2013, KUNDU et al. 2014, SAVINOV et al. 2015, VINEET et al. 2015) or, rarely, meshes (CABEZAS et al. 2015). All these works are variants of the same conceptual idea, to do the semantic labelling in 3D, such that it is inherently consistent across images, while at the same time enforcing class-specific priors rather than isotropic smoothing. (HÄNE et al. 2013) employ a non-metric regulariser, developed within the linear programming relaxation of a multi-label Markov random field (ZACH et al. 2014) in 3D voxel space. (KUNDU et al. 2014) is perhaps the closest work to ours. They also employ the octree data structure to save memory, but fix the local resolution at the beginning, based on an initial guess of the geometry (whereas we adaptively refine it). Also, like all other works mentioned above it uses only street-level imagery and models only scene parts visible from the streets, whereas we use aerial images and reconstruct the entire scene.

Before semantic (multi-label) 3D reconstruction, volumetric representations were already used for generic surface modelling (CURLESS & LEVOY 1996), where a distance field or indicator function is defined on the voxels. From that representation one can extract an explicit surface by finding the zero level set (e.g., LORENSEN & CLINE 1987, KAZHDAN et al. 2007). Many volumetric techniques work with regular voxels (CREMERS & KOLEV 2011, KOLEV et al. 2012, KOSTRIKOV et al. 2014). The data term, corresponding to a voxel's probability of lying behind the surface, is typically a signed distance generated from image correspondences (e.g., ZACH et al. 2007, ZACH 2008). Some authors even model a pixel's visibility along the complete ray (LIU & COOPER 2010, SAVINOV et al. 2015), which however leads to higher-order potentials over all voxels intersected by each ray, such that these methods do not scale to larger datasets. Instead of the regular voxels (LABATUT et al. 2007, JANCOSEK & PAJDLA 2011) tessellate the space by Delaunay tetrahedralisation of the initial multi-view point cloud. Tetrahedrons are labelled as empty or occupied, and triangles on the interface between the two labels form the object surface.

Algorithms for surface reconstruction from point clouds, like the Poisson method (KAZHDAN et al. 2006), also use octrees; and in some cases also multi-grid solvers (BOLITHO et al. 2007). This is similar in spirit to our adaptive multi-scale approach, but the least-squares nature of Poisson reconstruction is susceptible to outliers. Our model allows for robust error functions, at the cost of a more complicated optimisation. Also, our octree structure is not determined once and for all by the input data, but refined adaptively. A different view on our model is to see it as a coarse-to-fine reconstruction on a volumetric pyramid (ZACH et al. 2007, ZACH 2008), in which the refinement is applied selectively (e.g., HORNUNG & KOBBELT 2006).

3 Method

To simplify the technical description, we first describe the basic model with a regular voxel grid, and then extend it to the irregular, adaptive scheme. Throughout, the description stays on a conceptual level, for mathematical details the interested reader is referred to (BLAHÁ et al. 2016).

3.1 Basic Model

Let the region of interest Ω be discretised into regular, equally sized voxels $s \in \Omega$. In this discrete representation, joint geometric and semantic reconstruction can be cast as a labelling problem: at each voxel, determine the most likely semantic class label. By including a class *freespace*, one at the same time estimates also which voxels do not belong to any object, and thus implicitly reconstructs the 3D surfaces that separate objects from freespace. At each voxel an indicator vector $x^i \in [0,1]$ is stored, which is 1 for the assigned class and 0 for all other classes. The trick of (HÄNE et al. 2013) is to additionally store pseudo-marginals x^{ij} for each pair of classes and each grid direction, such that class transitions and their orientations are made explicit. Finding the best labelling then amount to minimising the energy function

$$E(x) = \sum_{s \in \Omega} \left(\sum_{classes\ i} \rho_s^i x^i + \sum_{class\_pairs\ i,j} \varphi^{ij}\left(x_s^{ij} - x_s^{ji}\right) \right),$$

subject to appropriate constraints which ensure that the variables are non-negative, that they sum to 1 appropriately, and that the x^i and x^{ij} are consistent. The ρ^i are conventional data terms that encode the likelihood of different labels at voxel *s*. The convex and 1-homogeneous functions φ^{ij} encode the individual a-priori likelihoods of different class transitions, taking into account the orientation $(x^{ij} - x^{ji}) \in [-1,1]^3$ of the boundary surface. The energy can be seen as a generalised form of the standard Markov Random Field energy (in its linear programming relaxation).

The **data term** ρ^i combines depth values and semantic class probabilities observed in the images. It favours configurations where the transition from freespace to another class occurs at the predicted depth along a pixel's viewing ray, and penalises deviations from that depth with a truncated linear penalty. At the same time, ρ^i maps the class likelihoods from the images into the 3D volume, by applying them at a voxel along the ray, slightly behind the predicted surface. The data cost in a voxel is computed by summing over all rays that pass through it. We point out that truncating the data cost means that we require the volume to be empty only in a limited interval in front of the predicted depth, rather than along the entire ray. This is clearly an approximation, but has the advantage that the data evidence can be encoded as a single unary term per voxel. In contrast, taking into account the full length of the viewing rays would lead to a higher-order potential that links all voxels along the ray, for each single pixel (LIU & COOPER 2010, KUNDU et al. 2014, SAVINOV et al. 2015). An efficient treatment of such potentials remains future work.

The class-specific **priors** φ^{ij} penalise class transitions, but other than Potts-type smoothing they account for the type of transition and the corresponding surface orientation. The prior is modelled as a sum of two terms, an isotropic part that encodes the likelihood (frequency) of the transition from class *i* to class *j*; and an anisotropic part that increases the penalty if the transition occurs in an improbable direction (e.g., an overhanging building wall). A convenient way to encode such a non-metric (w.r.t. the label space) and direction-dependent regulariser is through the indicator function of an appropriately chosen convex set, the so-called *Wulff* shape (ZACH et al. 2014).

3.2 Adaptive Multi-scale Extension

The basic model described so far links semantic interpretation and shape reconstruction in a principled manner, but it needs a lot of memory and processing power. Instead of exhaustively storing variables for the finest grid resolution, we reduce the set of unknowns by storing voxels (and their indicator variables) at an adaptive resolution: finer close to the surface boundaries, coarse further away from them. Any voxel stores only one set of pseudo-marginals x^{ij}, regardless of its size, so that computational resources are saved for non-expanded voxels. For the new, adaptive discretisation Ω^l of the space, we seek an energy E_l that approximates the original energy E (in which all voxels have the finest resolution) as tight as possible.

It turns one can see the E_l over the reduced set of unknowns (induced by the adaptive discretisation) as a constrained version of the original energy E, subject to additional equality

constraints. To ensure that the two energies are identical, the regularisation φ^{ij} must change as a function of the voxel size, respectively refinement level. Intuitively speaking, one has to compensate for the fact that, as the refinement proceeds, the transition penalties are increasingly "concentrated" in a smaller fraction of the overall volume and on fewer surfaces of the boundary voxels. The transition cost Φ_l^{ij} at a certain refinement level is a weighted sum of the "virtual transition costs" that would apply at the highest voxel resolution. The overall energy now reads

$$E_l(x) = \sum_{s \in \Omega^l} \left(\sum_{classes\ i} \rho_s^i x^i + \sum_{class\_pairs\ i,j} \Phi_l^{ij} \left(x_s^{ij} - x_s^{ji} \right) \right).$$

It sums the data costs over the voxels $s \in \Omega^l$ of the current refinement stage, and applies the corresponding level-dependent transition penalties Φ_l^{ij}. Non-negativity and normalisation constraints remain the same, consistency constraints must take into account that a voxel now may, at any of its faces, meet a single voxel of the same size, multiple smaller voxels, or part of a bigger voxel. For technical details and equations please see (BLAHA et al. 2016).

In our algorithm the scene is initially reconstructed on a very coarse grid and adaptively subdivided only close to the (putative) class boundaries. The refined variable set for the new, smaller voxels is initialised from the intermediate solution at the previous level. The alternation between energy minimisation and refinement is repeated, until the smallest voxels have reached the final target resolution. Note, due to the implicit representation the surface topology can change, e.g., a narrow street might open between two previously connected buildings.

The last missing piece is a criterion to decide which voxels to refine for the next round of energy minimisation. We simply refine all voxels that, at the current resolution, would be assigned to a different class than any of their neighbours. Moreover, we limit the resolution difference between adjacent voxels to at most 1, so splitting a voxel may trigger additional splits. In every iteration of the adaptive refinement, the energy function is, by construction, convex and can be minimised with standard tools. We convert it to primal-dual form with the help of Lagrange multipliers for the constraints and use the numerical scheme of (CHAMBOLLE & POCK 2011).

4 Experiments

As test dataset we use a block of 510 oblique and nadir images (102 exposures, Maltese cross configuration) from the city of Enschede, Netherlands. After orientation, per-image evidence for the depth is generated with semi-global matching (HIRSCHMÜLLER 2008), and class likelihoods are estimated with a MultiBoost classifier (BENBOUZID et al. 2012), using both RGB intensities and local shape features (CHEHATA et al. 2009). Figure 1 illustrates these pre-processing steps.

We impose two types of class-specific priors (Wulff shapes). One prefers horizontal surfaces with upward-pointing normal, and gradually increases the cost as the normal vector is tilted. This prior describes the boundaries *ground–freespace*, *ground–building*, *ground–vegetation*, *building–roof*, *roof–freespace*. The second prior has a strong preference for vertical transitions

building–freespace and *building–vegetation*. The strength of the priors (tolerance to deviations from the preferred orientation) is set individually per label pair.

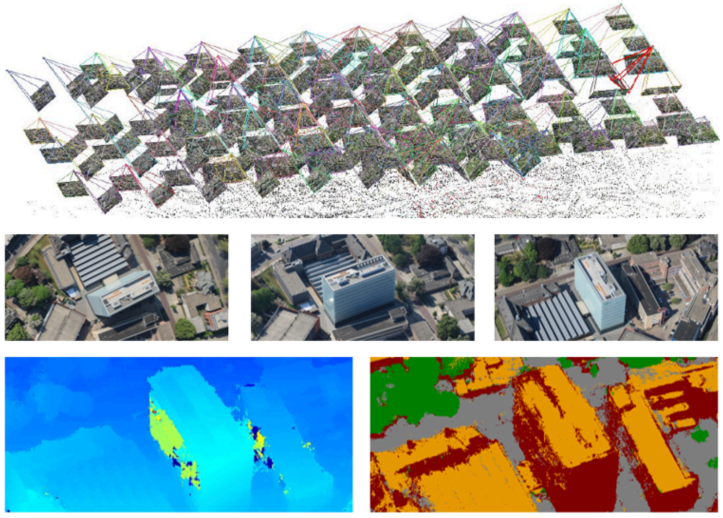

Figure 1: Input data: Image orientations *(top)*, depth maps *(bottom left)* and class likelihoods *(bottom right, illustration shows the most likely labels per pixel)*.

In a first experiment, we compare the adaptive model to a fixed voxel grid with the same target resolution. Unfortunately, no 3D ground truth was available. As a pragmatic compromise, we evaluate semantic correctness in image space: Semantic segmentations are hand-labelled for two representative images. Both reconstructions are back-projected into those images and quantitatively compared on a per-pixel basis. Overall the differences between adaptive and non-adaptive reconstruction are tiny (<0.7 percent points) and mostly due to aliasing, see Figure 2. We conclude that the hierarchical scheme does not incur any loss of accuracy. In this context we point out a remaining systematic error of the current model (with or without adaptive refinement). The approximate data term demands that behind a surface a few voxels are occupied along each ray, which leads to fattening at thin surfaces and silhouette edges. The effect is best visible at the transition from *building* to *roof*. It could be mitigated by correctly modelling visibility along entire rays (SAVINOV et al. 2015), with much higher memory consumption.

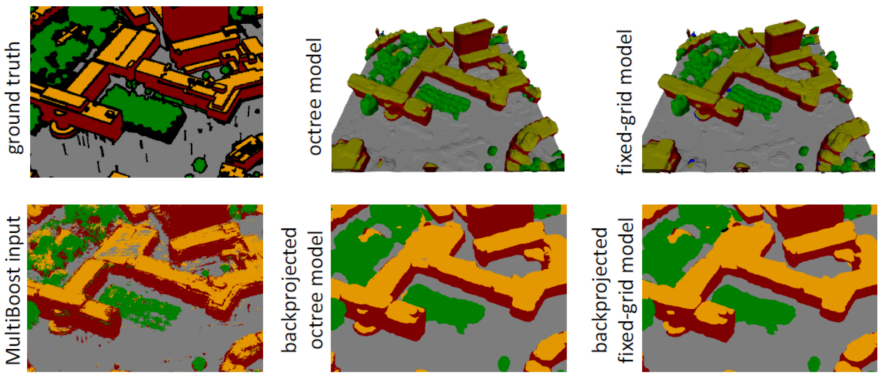

Figure 2: Comparison of classification accuracy in 2D.

The benefits of joint reconstruction are bigger in the other direction, when semantic knowledge improves the surface geometry in regions where matching fails, see examples in Figure 3. Unfortunately, we do not have reference data to quantify this improvement.

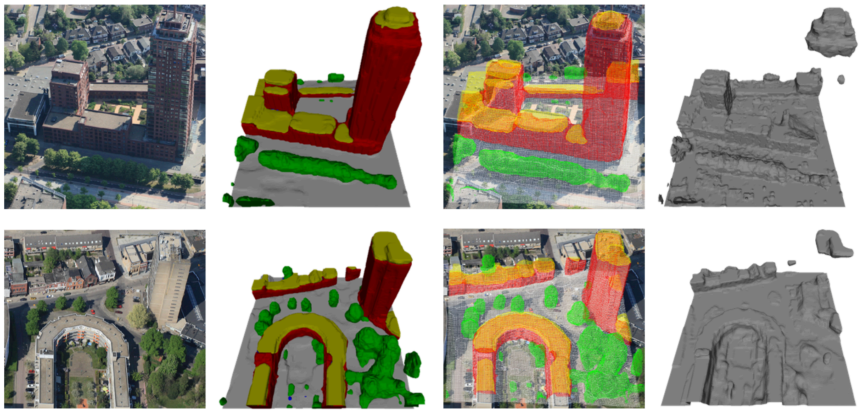

Figure 3: Reconstruction with and without class-specific priors. Left to right: example image; semantic reconstruction; semantic reconstruction back-projected to the image; and generic volumetric surface reconstruction with the same input (ZACH 2008).

We go on to measure how much memory and computation time can be saved with the adaptive scheme. All experiments are run on a machine with 64 GB of RAM and a hexa-core *Intel i7* CPU. Due to the huge memory consumption of the baseline (non-adaptive) method we have to limit this comparison to small sub-regions of the data. To ensure the comparison is fair we only voxelise a tight bounding box around the data. The tight box is a lot smaller for Enschede, because the terrain is flat. In mountainous cities the savings would be even higher. The top level of the octree has voxel size 13.5 m^3, the final resolution after five refinements is 0.4 m^3. In all tests the adaptive scheme saves ca. 95%, see Table 1. For the two smaller scenes 3 and 4 we have run another refinement down to 0.2 m^3 voxel size. At that resolution we can no longer run the baseline, since it would require > 108 GB of memory, ca. 35× more than the adaptive scheme. Figure 4 illustrates the evolution of the adaptive refinement. At the top, one can see how the accuracy and detail of the reconstruction gradually improves. The bottom row displays the voxel size in a vertical slice, colour-coded from blue (13.5 m^3) to yellow (0.2 m^3).

Table 1: Adaptive vs. non-adaptive volumetric reconstruction on two different test scenes.

	runtime # 0.4m [sec]		*memory @ 0.4m* [GB]		*memory @ 0.2m* [GB]	
	scene 1	*scene 2*	*scene 1*	*scene 2*	*scene 1*	*scene 2*
voxel grid	91'982	92'893	13.6	13.6	108.5	108.5
octree	5'488	4'984	0.7	0.7	3.3	2.7
ratio	16.8	18.6	19.4	19.4	32.9	40.2

Finally, we come to the target of reconstructing a large urban area and process all the 510 images at once to cover the city centre of Enschede (ca. 3 km^2) with target voxel size 0.8 m^3 (1/2048 of the bounding volume), see Figures 5 and 6. For this reconstruction the adaptive scheme uses a modest 28 GB of memory and runs 40 hours on one PC. To process the same dataset without adaptive computation (2048×2048×128 voxels), one would need 434 GB of memory.

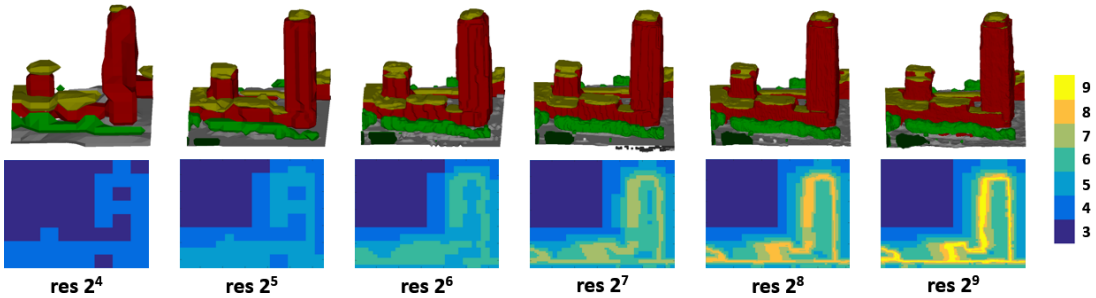

Figure 4: Evolution of adaptive reconstruction over five refinement steps.

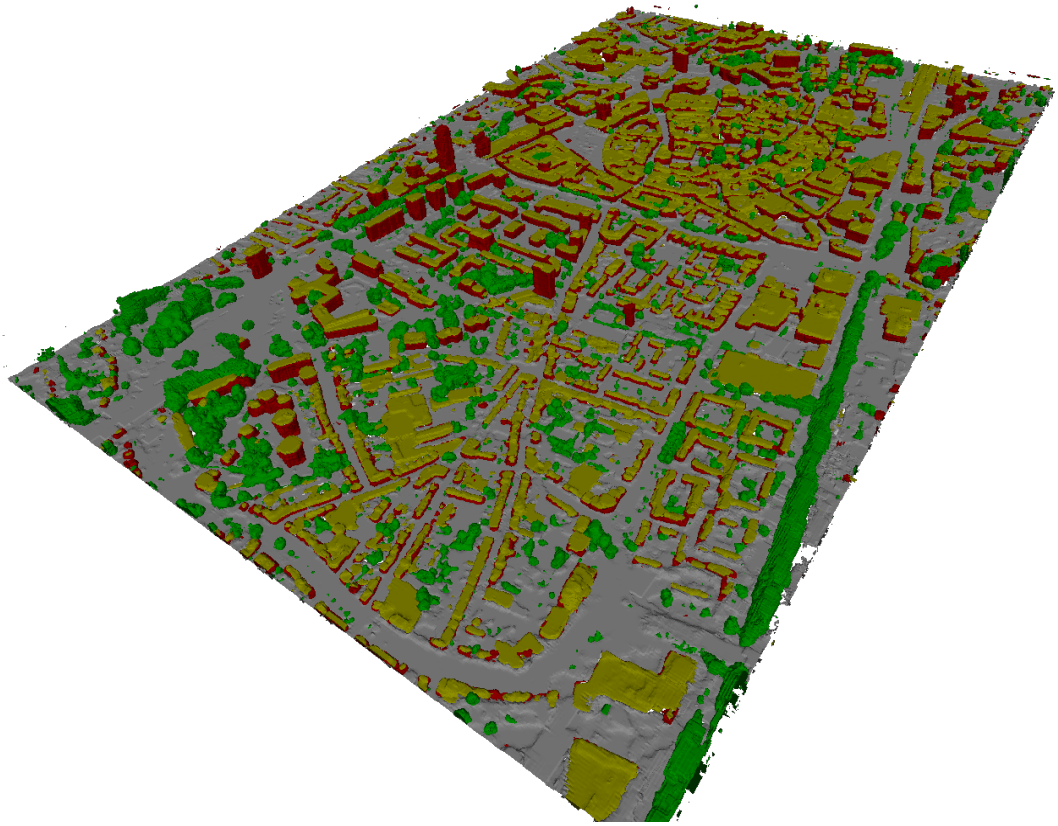

Figure 5: Semantic reconstruction of Enschede city centre.

5 Conclusion

We have described an adaptive multi-resolution framework to jointly infer the 3D geometry and a semantic segmentation of a scene from multi-view imagery, taking into account interactions between surface geometry and object type. The framework greatly improves the computational efficiency of integrated geometric/semantic 3D reconstruction, such that one can process scenes of realistic size and resolution. In future work we will investigate how to transfer the idea of adaptive spatial refinement to irregular space tessellations like the recently popular Delaunay tetrahedralisation, which by construction already adapt to the point distribution of the dataset.

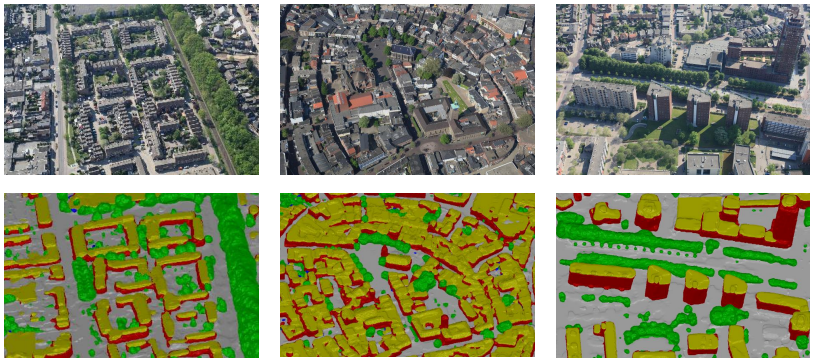

Figure 6: Visual comparison to aerial images with the same viewpoint.

Acknowledgements. We thank Christian Häne and Marc Pollefeys for source code and discussions. This work was supported by SNF grant 200021_157101.

6 Bibliography

BAO, Y., CHANDRAKER, M., LIN, Y. & SAVARESE, S., 2013. Dense object reconstruction using semantic priors. Computer Vision and Pattern Recognition (CVPR).

BENBOUZID, D., BUSA-FEKETE, R., CASAGRANDE, N., COLLIN, F.-D. & AND KÉGL, B., 2012. MULTIBOOST: a multi-purpose boosting package. Journal of Machine Learning Research, **13**(1), 549-553.

BLAHÁ, M., VOGEL, C., RICHARD, A., WEGNER, J., POCK, T. & SCHINDLER, K., 2016. Large-scale semantic 3d reconstruction: an adaptive multi-resolution model for multi-class volumetric labeling. Computer Vision and Pattern Recognition (CVPR).

BOLITHO, M., KAZHDAN, M., BURNS, R. & HOPPE, H., 2007. Multilevel streaming for out-of-core surface reconstruction. Eurographics.

CABEZAS, R., STRAUB, J. & FISHER III, J. W., 2015. Semantically-aware aerial reconstruction from multi-modal data. International Conference on Computer Vision (ICCV).

CHAMBOLLE, A. & POCK, T., 2011. A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of Mathematical Imaging and Vision, **40**(1), 120-145.

CHEHATA, N., GUO, L. & MALLET, C., 2009. Airborne LiDAR feature selection for urban classification using random forests. ISPRS Archives, **38**(3).

CREMERS, D. & KOLEV, K., 2011. Multiview stereo and silhouette consistency via convex functionals over convex domains. IEEE Transactions on Pattern Analysis and Machine Intelligence, **33**(6), 1161-1174.

CURLESS, B. & LEVOY, M., 1996. A volumetric method for building complex models from range images. ACM SIGGRAPH.

HÄNE, C., ZACH, C., COHEN, A., ANGST, R. & POLLEFEYS, M., 2013. Joint 3d scene reconstruction and class segmentation. Computer Vision and Pattern Recognition (CVPR).

HIRSCHMÜLLER, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence, **30**(2), 328-341.

HORNUNG, A. & KOBBELT, L., 2006. Hierarchical volumetric multiview stereo reconstruction of

manifold surfaces based on dual graph embedding. Computer Vision and Pattern Recognition (CVPR).

JANCOSEK, M. & PAJDLA, T., 2011. Multi-view reconstruction pre- serving weakly-supported surfaces. Computer Vision and Pattern Recognition (CVPR).

KAZHDAN, M., BOLITHO, M. & HOPPE, H., 2006. Poisson surface reconstruction. Eurographics.

KAZHDAN, M., KLEIN, A., DALAL, K. & HOPPE, H., 2007. Unconstrained isosurface extraction on arbitrary octrees. Eurographics.

KOLEV, K., BROX, T. & CREMERS, D., 2012. Fast joint estimation of silhouettes and dense 3D geometry from multiple images. IEEE Transactions on Pattern Analysis and Machine Intelligence, **34**(3), 493-505.

KOSTRIKOV, I., HORBERT, E. & LEIBE, B., 2014. Probabilistic labeling cost for high-accuracy multi-view reconstruction. Computer Vision and Pattern Recognition (CVPR).

KUNDU, A., LI, Y., DELLAERT, F., LI, F. & REHG, J., 2014. Joint semantic segmentation and 3d reconstruction from monocular video. European Conference on Computer Vision (ECCV).

LABATUT, P., PONS, J.-P. & KERIVEN, R., 2007. Efficient Multi-View Reconstruction of Large-Scale Scenes using Interest Points, Delaunay Triangulation and Graph Cuts. International Conference on Computer Vision (ICCV).

LADICKÝ, L., STURGESS, P., RUSSELL, C., SENGUPTA, S., BASTANLAR, Y., CLOCKSIN, W. & TORR, P., 2010. Joint optimisation for object class segmentation and dense stereo reconstruction. British Machine Vision Conference (BMVC).

LAFARGE, F. & MALLET, C., 2012. Creating large-scale city models from 3D-point clouds: a robust approach with hybrid representation. International Journal of Computer Vision, **99**(1), 69-85.

LIU, S. & COOPER, D. B., 2010. Ray Markov random fields for image-based 3d modeling: Model and efficient inference. Computer Vision and Pattern Recognition (CVPR).

LORENSEN, W. E. & CLINE, H.E., 1987. Marching cubes: A high resolution 3d surface construction algorithm. ACM SIGGRAPH.

SAVINOV, N., LADICKÝ, L., HÄNE, C. & POLLEFEYS, M., 2015. Discrete optimization of ray potentials for semantic 3d reconstruction. Computer Vision and Pattern Recognition (CVPR).

VINEET, V., et al., 2015. Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction. International Conference on Robotics and Automation (ICRA).

ZACH, C., 2008. Fast and high quality fusion of depth maps. International Conference on 3D Vision (3DV).

ZACH, C., HÄNE, C. & POLLEFEYS, M., 2014. What is optimized in convex relaxations for multilabel problems: Connecting discrete and continuously inspired MAP inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, **36**(1), 157-170.

ZACH, C., POCK, T. & BISCHOF, H., 2007. A globally optimal algorithm for robust TV-L1 range image integration. International Conference on Computer Vision (ICCV).